THE DESIGN AND REQUIREMENTS EVOLUTION OF A SPEECH RECOGNITON TECHNOLOGY FOR TACTICAL APPLICATIONS AND ENVIRONMENTS

Mr. Lockwood Reed
US Army RDECOM CERDEC
Fort Monmouth, NJ 07703

ABSTRACT

This paper presents a discussion of the unique and specialized requirements for a militarized speech recognizer. It also presents the tactical advantages of militarized speech recognition technology, as it could be applied in several Command and Control (C2) applications and environments. Additionally, this paper will present the results of a comparison study, which was performed between a custom military speech recognition technology and various manual input modalities, including keyboard and trackball, for activating a selected C2 application. The results of this paper demonstrate a clear superiority of continuous speech recognition over discrete speech recognition in both metrics, and a tradeoff of task execution speed for error rate for continuous speech recognition verses manual input

1. BACKGROUND

Initial research into speech recognition technology was to utilize the extensive analysis capabilities of the Command and Control Directorate Audio Laboratory to evaluate the effectiveness of Commercial Off The Shelf (COTS) speech recognition technology for command and control applications in the Army Aviation environment. Once all the available COTS speech recognition technology was evaluated and the deficiencies identified, specifications and requirements could be written. One critical requirement was for operation in high noise ranging from 103-107dBA, for the Blackhawk type helicopter and 110dBA for the Apache helicopter, the two most commonly used aircraft. Other aircraft, such as the CH-47 Chinook and the Heavy-lift Helicopter produced sound levels of 115dBA and 123dBA respectively. The unassisted COTS technology failed to perform reliably at sound levels as low as 80dBA, which presented quite a technological challenge. Initially we simply had test subjects enroll in the target environment and recognition accuracy jumped from the low 70 percentile to over 95%. These results were reported at the subsequent Interactive Speech Technical Advisory Committee (ISTAC) meeting, and various member organizations confirmed the results. Early on we knew that this approach had its limitations: first and foremost it was too stressful on the test subjects to enroll in high noise environments; second the recognition technology became too dependent upon the environment. The Air Force group at the Rome Air Development Center (RADC) experimented with reducing the level of the enrollment environment[1], while normal operation was evaluated at normal environmental levels. Helicopter environments are relatively stationary. The environment does not spectrally vary enough during various flight profiles to affect recognition performance. However, the noise environment of military track vehicles is not nearly as stationary as the helicopter environment, and these variations must be taken into consideration. For our own work we wanted to eliminate the user from having to enroll in the environment altogether, and we wanted to eliminate environmental dependency. Our approach to eliminating the noise during the enrollment session (at least as far as the test subject was concerned) was to electrically mix environmental noise into the test subject input stream to the target speech recognizer. The RADC work indicated that the precise signal-to-noise ratio should not be too critical, and indeed, through electrical mixing, we obtained results comparable to actual acoustical mixing of signal and noise (i.e. test subject in the environment). This also assuaged fears that the Lombard-Effect[2] would affect the results. To achieve environment independence, the recognition technology would need to adapt to environmental changes on the fly, during operational use. Our technology provides reliable performance in noise levels up to 115dBA, does not require user enrollment in the target environment, will adapt to changing noise environments, and can be configured to permit the user to whisper, speak normally or shout commands.

2. DISCUSSION

2.1 Noise Processing

It became evident early in the evolution of our speech recognition technology that while overall sound level adversely affected recognition performance, the effect was not linear with increasing sound level. When the sound reached a certain critical level for a given recognition system, performance degraded rapidly and then failed to be usable completely. The recognition systems were found to be far more sensitive to changes in the spectral content of the noise environment, than simply to changes in the sound intensity. It is believed that this relationship of performance to the sound characteristics accounts for the improved performance obtained by the additive techniques over subtractive techniques. Subsequent dynamic versions of the subtractive technique were experimented with, but the dynamic additive

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
00 DEC 2004	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The Design And Requirements Evolution Of A Speech Recogniton Technology For Tactical Applications And Environments	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **US Army RDECOM CERDEC Fort Monmouth, NJ 07703**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001736, Proceedings for the Army Science Conference (24th) Held on 29 November - 2 December 2005 in Orlando, Florida.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**2**	

technique maintained a significant performance enhancement.

2.2 Gain Management

The next technical hurdle to overcome was the sensitivity of the recognizer technology to speaker dynamics and microphone placement. The implementation employed in our recognizer is deterministic. It is an integrated component of the recognition algorithm. Future implementations of the recognizer will employ technology, which will accommodate the full range of speaker dynamics, eliminating the need for Automatic Gain Control (AGC).

2.3 Whispered/Shouted Speech

Our current technology permits multiple template sets to be active simultaneously. Therefore normal and whispered template sets can exist side by side, allowing the user either mode of operation. Additionally, shouted speech is handled in a similar manner. While this can provide an impressive demonstration, it is not the ultimate solution we envision. We have devised a far more sophisticated approach, which does not require alternate template sets. Unfortunately, as with most technological advances, the implementation of the more sophisticated approach is awaiting additional investment, as it will require modification and addition to the current architecture – but it is imminently achievable.

2.4 Speech Recognition System (SRS) Testing Results

The same tasks took 83% longer to perform by isolated word recognition as compared to continuous speech; manual mode took 92% longer to perform the same task as compared to continuous speech; and isolated word recognition was only 6% faster then manual mode. The results are based on 913 manual operations, 850 isolated utterances (words) and 904 continuous utterances (words). In addition to the quantifiable information collected, the test subjects were asked to provide subjective scores to the following questions: On a scale of 1 to 5, where 1 is easiest and 5 is hardest: How easy was it to use (continuous recognition, isolated recognition, and manual entry)? The average score for the 18 participants breaks down as follows: continuous (1.3), isolated (2.5), and manual (2.3). The participants were also asked a related question, which attempted to ascertain how confident the subject felt with the respective entry modalities: On a scale of 1 to 5, where 5 is the most comfortable and 1 is the least: how comfortable did you feel with (continuous recognition, isolated recognition, and manual entry)? The average score for the 18 participants breaks down as follows: continuous (3.7), isolated (2.5), and manual (3.1). It is interesting to note that continuous speech recognition was on par with manual entry in this static test, which favored manual input. Experiments performed by other agencies, including the Air Force[3], have shown that speech is relatively insensitive to movement, as compared to manual operation. Given the certain increase in manual entry error rate for a moving vehicle, it is not unreasonable to predict that the scores would shift, further favoring continuous speech recognition over manual entry in a dynamic environment.

REFERENCES

[1]. "Speech Enhancement for Improved Recognition" By Michael Heffron, Rome Air Development Center, Minutes of the Voice Interactive Systems SUBTAG, 7 March 1983.

[2] J.C. Junqua and Yolanda Angelade, "Acoustic and perceptual studies of Lombard speech: application to isolated-words automatic speech recognition", Proc. ICASSP 90, 841-844, Albuquerque, NM, April 1990.

[3] "AFTI/F-16 Voice Command Systems: Status Report" By Major Stephen F. Gray,Edwards Air Force Base, Minutes of the Voice Interactive Systems SUBTAG, 7 March 1983.

CONCLUSIONS

Since 9/11 the world has become quite a different place. Our military face new and formidable challenges. The rapid pace and success of "Iraqi Freedom" lends credence to the effectiveness of highly mobile forces, and confirms the need for command-on-the-move technologies. In addition, as evidenced in Iraq and Afghanistan, the battles will be fought from street-to-street, house-to-house, room-to-room and cave-to-cave. A soldier can become a statistic "in a heartbeat" if he/she is even momentarily distracted from maintaining a hands-on-weapon, eyes-alert posture, through having to manually interact with some tactical system. Additionally, soldiers will need the capability to interact with these tactical systems, while running and firing. Speech recognition technology has demonstrated the capability to provide hands-free, eyes-free activation of tactical systems. Further it has also demonstrated its ability to operate under harsh and high noise battlefield environments. While no system can claim to operate flawlessly, under all battlefield conditions, there is sufficient evidence to conclude that the current state-of-the-art tactical SRS technology can fulfill 90% of the current needs, to enable faster more intuitive Soldier/machine interaction, resulting in increased task accuracy, reduced task time, and ultimately yielding greater survivability and lethality. Additional, technology has been identified which, for a minimum investment, can improve the current technology to pickup most of the remaining 10%.